基于随机蛙跳和支持向量机的牛乳收购分级模型构建

肖仕杰[1]，王巧华[1,2*]，李春芳[3,4]，赵利梅[4]，刘鑫雅[4]，卢士宇[4]，张淑君[3*]

（1. 华中农业大学 工学院，湖北武汉 430070；2. 农业农村部长江中下游农业装备重点实验室，湖北武汉 430070；3. 华中农业大学农业动物遗传育种与繁殖教育部重点实验室，湖北武汉 430070；4. 河北省畜牧业协会，河北石家庄 050031）

摘　要：蛋白质、脂肪含量和体细胞数量作为牛乳收购中的重要参考指标，决定了牛乳的品质和价格。为批量准确地对牛乳品质进行分级，提高乳企的生产效率，本研究以3216份荷斯坦牛牛乳样本为研究对象，应用中红外光谱技术实现对收购过程中4种不同品质牛乳的检测分级。利用一阶导数和一阶差分对光谱进行预处理，并结合竞争性自适应重加权算法（Competitive Adaptive Reweighted Sampling，CARS）和随机蛙跳算法（Shuffled Frog Leaping Algorithm，SFLA）筛选出能代表不同牛乳的有效特征变量，建立支持向量机（Support Vector Machine，SVM）模型。其中，利用网格搜索法（Grid Search，GS）、遗传算法（Genetic Algorithm，GA）和粒子群算法（Particle Swarm Optimization，PSO）对SVM模型的关键参数——惩罚参数c和核函数参数g进行优化。结果表明，SFLA算法总体上优于CARS算法，PSO优化SVM模型的效果最佳。一阶差分预处理后，利用SFLA算法筛选特征变量建立的PSO-SVM模型，训练集准确率、测试集准确率和曲线下面积（Area Under Curve，AUC）分别为97.8%、95.6%和0.96489。该模型具有较高的准确率，在牛乳产业中具有实际应用价值。

关键词：中红外光谱；牛乳；收购分级；随机蛙跳；支持向量机

中图分类号：S37；O657.33　　　**文献标志码**：A　　　**文章编号**：202107-SA003

引用格式：肖仕杰, 王巧华, 李春芳, 赵利梅, 刘鑫雅, 卢士宇, 张淑君. 基于随机蛙跳和支持向量机的牛乳收购分级模型构建[J]. 智慧农业(中英文), 2021, 3(4): 77-85.

XIAO Shijie, WANG Qiaohua, LI Chunfang, ZHAO Limei, LIU Xinya, LU Shiyu, ZHANG Shujun. Construction of milk purchase classification model based on shuffled frog leaping algorithm and support vector machine[J]. Smart Agriculture, 2021, 3(4): 77-85. (in Chinese with English abstract)

1　引　言

蛋白质和脂肪是牛乳的重要营养组成成分，是决定牛乳品质和价格的关键因素。乳脂和乳蛋白含量一直作为牛乳的收购参考指标[1]。此外，体细胞数（Somatic Cell Count，SCC）会直接影响牛乳中的蛋白质和脂肪含量以及奶牛的产奶量。现如今，SCC已被乳品行业作为牛乳收购标准之一[2]。乳脂、乳蛋白和体细胞数在欧美被第三方实验室广泛用于按质论价[3]。1 mL牛乳中

收稿日期：2021-07-05　修订日期：2021-08-08

基金项目：欧盟FP7构架项目（FP7-KBBE-2013-7-613689）；国家重点研发计划（2017YFD0502002）

作者简介：肖仕杰（1993—），男，硕士，研究方向为智能化检测与测控技术。E-mail：1175760869@qq.com。

*通信作者：1. 王巧华（1970—），女，博士，教授，研究方向为农产品智能化检测技术与装备。电话：027-87282120。E-mail：wqh@mail.hzau.edu.cn；2. 张淑君（1966—），女，博士，教授，研究方向为动物遗传育种与繁殖。电话：027-87282091。E-mail：sjxiaozhang@mail.hzau.edu.cn。

SCC不高于20万个时，奶牛身体为健康状态[4]；当超过此界限，SCC数量递增的同时，牛乳的品质以及奶牛的产奶量均会下降；1 mL牛乳中SCC高于50万个时，奶牛有很大机率感染亚临床乳房炎（乳腺炎），牛乳品质进一步降低[2]；若1 mL牛乳中SCC高于100万个，奶牛很有可能患有临床乳房炎[5]。

近年来，中红外光谱（Mid-infrared Spectroscopy，MIRS）被广泛用于牛乳中各成分的无损检测[6-9]，因此MIRS技术为牛乳收购中鉴别不同等级牛乳提供了一种有效手段。然而，MIRS的波长范围广，在包含丰富有效信息的同时，也包含很多冗余信息和背景噪声，对模型造成干扰。因此，分析并揭示中红外光谱的响应规律，筛选出最能代表不同等级牛乳的差异波段，对简化MIRS模型、提高模型精度和效率具有重要意义。

筛选变量的方法可分为三大类：变量优化选择法、变量区间选择法以及变量信息选择法。变量优化选择法通过创造一个合适的目标函数以寻找最优变量组合，主要包括遗传算法（Genetic Algorithm，GA）[10]、粒子群算法（Particle Swarm Optimization，PSO）[11]、模拟退火算法（Simulated Annealing Algorithm，SAA）[12,13]等。变量区间选择法通过筛选多个光谱区间作为有效变量区间组合，区间数目的选择直接影响模型的结果，包括间隔偏最小二乘法（Interval Partial Least Squares，iPLS）[14]、前向间隔偏最小二乘法（Forward Interval PLS，FiPLS）[15]、后向间隔偏最小二乘法（Backward Interval PLS，BiPLS）[16]和移动窗口最小二乘法（Moving Window Partial Least Squares，MWPLS）[17,18]。变量信息选择法将信息变量作为描述变量在模型中所起作用大小的指示变量，其中竞争性自适应重加权算法（Competitive Adaptive Reweighted Sampling，CARS）[19]、无信息变量消除法（Elimination of Uninformative Variables，UVE）[20]和连续投影法（Successive Projections Algorithm，SPA）[19]是较为流行的几种算法。另外，一些新型的组合优化算法如随机蛙跳算法（Shuffled Frog Leaping Algorithm，SFLA）[21,22]被广泛应用于农业工程。此外，主成分分析（Principal Component Analysis，PCA）等算法也用于压缩数据，但一般不用来做直接的特征提取而是用来做特征矩阵的降维[23]。与CARS、SFLA算法等直接筛选代表性变量相比，主成分各个特征维度的含义具有一定的模糊性，不如原始样本特征的解释性强。

本研究以河北地区9个牧场的牛乳为研究对象，分析牛乳收购中不同等级牛乳的光谱特征，利用CARS算法和SFLA算法筛选特征变量，建立SVM模型，为MIRS技术在牛乳收购过程中提供支持。

2 材料与方法

2.1 试验材料

研究采用的3216份牛乳样本均来源于河北省9个牧场的荷斯坦牛。奶牛饲养于平均环境温度为10～29 ℃、相对湿度为45%～78%的可连续取水的牛棚。2019年11月～2020年10月（不包括2月）期间，从晨乳中收集样本。每个月采集一次样本，当天上午5:30开始采集，上午采完。使用全自动转盘挤奶设备逐头精确采集每头奶牛40 mL的牛乳，然后将牛乳放入从奶牛群体改良（Dairy Herd Improvement，DHI）检测实验室获得的全新特定采样瓶中，并依次编号。共采集3216份牛乳样本。为防止牛乳变质，在每个采样瓶中加入了专用防腐剂布罗波尔3.2～3.4 μL后，立即放入冰箱保存（4 ℃），并于第二天进行光谱采集。

2.2 试验方法

2.2.1 光谱采集、乳成分及SCC检测

设备：乳成分分析仪MilkoScanTM FT+（傅里叶变换中红外光谱仪FTIR），体细胞检测仪FossomaticTM7。

所有牛乳均在河北省DHI检测中心完成数据

采集。具体步骤为：将电热恒温水浴锅预热至（42±0.2）℃，将牛乳分批放入，加热15～20 min后摇晃均匀，使用MilkoScanTM FT+采集光谱以及蛋白质和脂肪含量测定。

此外，使用FossomaticTM7测定牛乳中的体细胞数。

2.2.2 收购分级标准

T/HLJNX 001-2018《黑龙江省食品安全团体标准》为黑龙江省乳制品企业牛乳收购和质量监督的参考依据，以此标准为参考标准，结合SCC进行分级。分级标准如表1所示。

表1 牛乳分级标准

Table 1 Standard of milk classification

级别	脂肪/%	蛋白质/%	SCC（10^4个/mL）
特级	≥3.6	≥3.2	≤20
一级	≥3.4且≤3.6	≥3.0且≤3.2	≤50
二级	≥3.2且≤3.4	≥2.8且≤3.0	≤50
低质量	<3.2	<2.8	≤100

2.2.3 样本划分

以表1为依据对牛乳分级。所有牛乳中，特级牛乳数量为940份，一级牛乳数量为826份，二级牛乳数量为537份，低质量牛乳数量为913份。按照约7:3的原则利用随机划分RS（Random Selection）算法将样本集划分为训练集和测试集。样本集的划分情况如表2所示。

表2 牛乳样本集的划分

Table 2 Division of milk sample sets

样本集	训练集/份	测试集/份
特级	658	282
一级	578	248
二级	376	161
低质量	640	273

2.3 数据处理

2.3.1 特征变量筛选

CARS算法以降低无信息变量为出发点，模型运行过程中，以PLS回归系数为衡量标准，根据交叉验证均方根误差（Root Mean Square Error of Cross-Validation，RMSECV）对应的位置选择最优的子集代表特级、一级、二级和低质量牛乳中红外光谱差异的特征变量组合。

SFLA算法将全局搜索性能良好的粒子群算法和局部搜索能力较强的元算法进行结合，从而可以获得强大的寻优能力。

2.3.2 基于参数寻优的支持向量机模型

支持向量机（Support Vector Machine，SVM）[24]是基于机器学习方法的强大多元技术，由Vapnik和Burges首次引入[25,26]。简单来说，SVM利用核技巧将输入向量映射到更高维的特征空间中，然后构造最大边距分离超平面进行特级、一级、二级和低质量牛乳的分类。在本研究中，使用径向基函数（Radial Basis Function，RBF）构建模型，利用网格搜索法（Grid Search，GS）、GA和PSO对RBF核函数的两个重要参数惩罚参数c和核函数参数g进行优化，分别建立GS-SVM、GA-SVM和PSO-SVM模型。

2.3.3 模型评估

利用准确率作为模型的主要评价指标，训练集准确率与测试集准确率越高且两者越接近，表明模型的精度高，可靠性好。

$$\text{准确率} = \frac{\text{预测正确的样本}}{\text{总样本}} \times 100\% \qquad (1)$$

3 结果与讨论

3.1 不同牛乳的光谱分析

特级、一级、二级和低质量牛乳在MIRS范围内的原始吸收曲线如图1。可以看出，牛乳的光谱曲线严重重叠，由于水的干扰，1597～1712 cm^{-1}和3024～3 680 cm^{-1}左右的区域信噪比低[27]，无法用于建模。不同牛乳的平均光谱曲线走向趋势相似（图2），表明它们的内部化学成分基本一致，但同时它们的光谱又存在差异，表明4类牛乳的化学成分含量存在差异。其中，一级牛乳和二级牛乳的平均光谱十分接近，通过肉眼难以区分，特级和低质量牛乳则与它们存在一定差异。根据福斯公司提供的乳成分的吸收情况可知，1754 cm^{-1}左右的波峰主要与脂肪中C=O键的伸缩振动有关，2857 cm^{-1}左右的波峰主要与脂肪酸链中的饱和C-H键的伸缩振动有关，

1470 cm^{-1}左右的波峰主要与脂肪酸链中饱和C-H键的弯曲振动有关，1538 cm^{-1}主要与N-H键的弯曲振动有关。

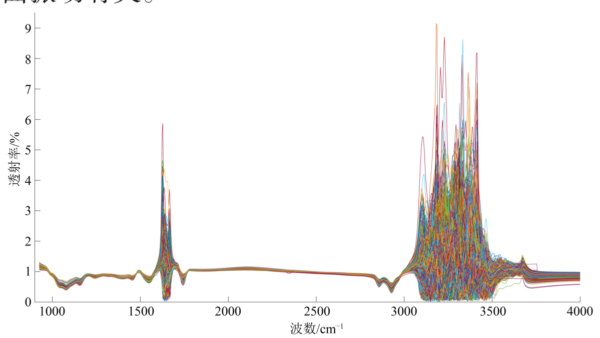


图1　特级牛乳、一级牛乳、二级牛乳和低质量牛乳的原始光谱

Fig. 1　Original spectra of premium milk，first-grade milk，second-grade milk and low quality raw milk

3.2　光谱预处理和特征变量选择

选择925～1597 cm^{-1}和1712～3024 cm^{-1}的敏感波段组合作为全光谱，分别利用一阶差分和一阶导数预处理。

预处理后的全光谱信息得到增强，但光谱维数过多，会导致SVM模型收敛速度慢，全光谱中还存在与牛乳分级不相关的变量，直接用于建模会对模型造成干扰。使用CARS算法、SFLA算法分别进一步提取有用变量，剔除无信息变量，找出能够代表特级、一级、二级和低质量牛乳的变量组合，简化模型，提高预测精度。

3.2.1　采样CARS算法筛选特征变量建立SVM模型

本研究利用5折交叉验证，将重采样率设置为0.8。将CARS的重采样次数分别设为50、100和200次，对比了不同重采样次数对SVM模型效果的影响，最终将重采样次数定

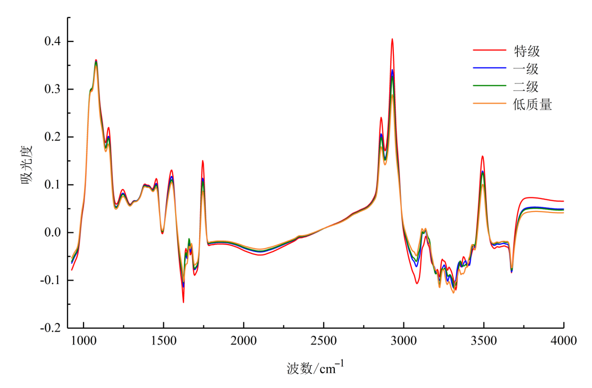
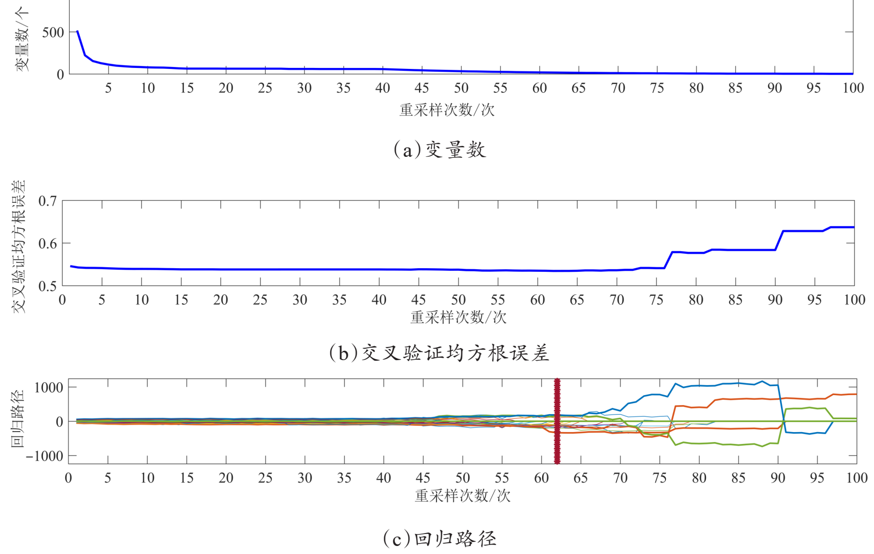


图2　特级牛乳、一级牛乳、二级牛乳和低质量牛乳的平均光谱

Fig. 2　Mean spectra of premium milk，first-grade milk，second-grade milk and low quality raw milk

为100次。以一阶导数预处理后的光谱数据为例阐述CARS算法进行变量选择的过程。图3（a）为被选取的特征变量数随着重采样运行次数的变化曲线。由图3（b）可知，在100次重采样中，当重采样次数为62时，对应最小交叉验证均方根误差值为0.5441，此时各变量的回归系数位于图3（c）中竖线位置，取得最优变量组合。

如表3所示，利用CARS算法对全光谱、一阶导数光谱和一阶差分光谱筛选的特征变量数分别为30、17和19，依次建立GS-SVM、GA-



（a）变量数

（b）交叉验证均方根误差

（c）回归路径

图3　竞争性自适应重加权算法筛选特征波长

Fig. 3　Screening characteristic wavelengths by competitive adaptive reweighted sampling algorithm

SVM和PSO-SVM分级模型。与全光谱数据相比，一阶导数处理后建立的分级模型预测性能均得到提升，而一阶差分处理后的分级模型预测性能均有所下降，三种模型均在一阶导数处理后获得最高的分级准确率，GS-SVM模型的效果优于GA-SVM模型和PSO-SVM模型，其训练集准确率为95.4%，测试集准确率为94.5%。

表3 CARS算法建立的SVM模型准确率结果

Table 3 Accuracy results of SVM models established using CARS algorithm

模型	预处理方法	特征变量数	训练集准确率/%	测试集准确率/%
GS-SVM	全光谱	30	95.2	93.8
	一阶导数	17	95.4	94.5
	一阶差分	19	95.4	93.6
GA-SVM	全光谱	30	95.6	93.8
	一阶导数	17	95.2	94.2
	一阶差分	19	95.0	93.6
PSO-SVM	全光谱	30	95.6	93.9
	一阶导数	17	95.1	94.2
	一阶差分	19	95.0	93.6

3.2.2 采样SFLA算法筛选特征变量建立SVM模型

本研究中设置SFLA运行次数N为10,000，最大潜在变量数A为6，抽样变量的初始数量Q为2。利用概率的大小作为变量筛选的评价指标，在图4中，横坐标代表每一维光谱变量的编号，纵坐标代表被选择的概率。波峰越高，表明变量被选中的可能性越大。以一阶差分处理后的515维光谱为例，将515个变量被选的概率排序，以0.1为阈值，最终得到位于图中虚线上方的146个最优变量组合。

如表4所示，利用SFLA算法对全光谱、一阶导数光谱和一阶差分光谱筛选的特征变量数分别为23、77和146，依次建立GS-SVM、GA-SVM和PSO-SVM分级模型。一阶导数和一阶差分处理后的分级模型性能均得到显著提高，三种模型在一阶差分处理后获得最高的分级准确率。其中，PSO-SVM模型的效果优于GS-SVM模型和GA-SVM模型，训练集准确率和测试集准确率分别为97.8%和95.6%。

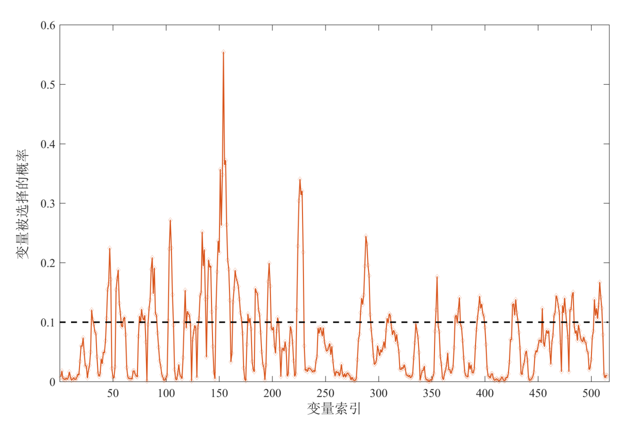


图4 随机蛙跳算法筛选特征波长

Fig. 4 Screening characteristic wavelengths by shuffled frog leaping algorithm

表4 SFLA算法建立的SVM模型准确率结果

Table 4 Accuracy results of SVM models established using SFLA algorithm

模型	预处理方法	特征变量数	训练集准确率/%	测试集准确率/%
GS-SVM	全光谱	23	92.9	90.8
	一阶导数	77	96.8	94.3
	一阶差分	146	96.5	95.5
GA-SVM	全光谱	23	92.2	90.2
	一阶导数	77	95.6	94.2
	一阶差分	146	97.7	95.3
PSO-SVM	全光谱	23	92.1	90.0
	一阶导数	77	95.9	94.3
	一阶差分	146	97.8	95.6

对比发现，未处理的全光谱结合CARS算法建立的模型测试集准确率为93.8%~93.9%，预处理后的全光谱结合CARS算法建立的模型测试集准确率为93.6%~94.5%（表3）。未处理的全光谱结合SFLA算法建立的模型测试集准确率为90.0%~90.8%，预处理后的全光谱结合SFLA算法建立的模型测试集准确率为94.2%~95.6%（表4）。无论是采用一阶导数还是一阶差分预处理，与未处理的全光谱相比，SFLA算法对模型性能的提升明显优于CARS算法。

SFLA算法建立的3种最佳模型均优于CARS算法建立的模型。其中，通过一阶导数-CARS算

法筛选的特征变量数仅占全光谱的3.29%，一阶差分-SFLA算法筛选的特征变量数占全光谱变量数的28.29%，因此，SFLA算法筛选的有效变量更多，更具代表性，模型的预测能力更强。

3.3 三种SVM模型对比

确定一阶差分-SFLA算法建立的模型效果最优后，对比不同的寻优算法对SVM模型的影响。同时，通过受试者工作特征（Receiver Operating Characteristic，ROC）曲线和ROC曲线下面积（Area Under Curve，AUC）[28]进一步评价三种SVM模型。ROC曲线可以反应分类器在某个阈值时对样本的识别能力，曲线越趋近坐标轴左上方位置，曲线下方与X轴围成的面积AUC越大，模型的性能越好。由图5可知，GS-SVM、GA-SVM、PSO-SVM模型ROC曲线的AUC分别为0.95786、0.95935和0.96489，PSO-SVM优于GS-SVM和GA-SVM模型。

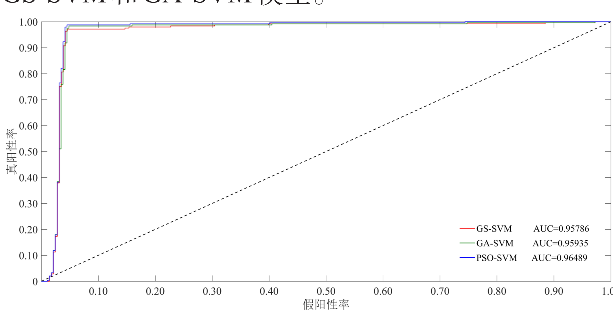


图5 三种SVM模型的ROC曲线

Fig. 5 ROC curves of three SVM models

对比发现，通过GS得到的c值较大，g值较小，而通过GA和PSO算法得到的c值较小，g值较大。其中，c与SVM算法对奇异点的重视程度有关，c值不宜过大或过小，否则会对模型精度造成影响；g与SVM算法的收敛速度有关，g越大，支持向量越少，模型收敛越快[29]。PSO寻优算法建立的SVM模型训练集准确率、测试集准确率和AUC值均优于GS和GA算法（表5）。对比训练时间，GS远大于GA和PSO算法，因此，综合考虑准确率、AUC值和训练时间，最终选择一阶差分-SFLA-PSO-SVM模型为最佳牛乳收购分级模型。

表5 三种参数寻优算法下的SVM的模型

Table 5 SVM models based on three parameter optimization algorithms

寻优算法	寻优参数		训练集准确率/%	测试集准确率/%	AUC	训练时间/s
	c	g				
GS	1024	84.4485	96.5	95.5	0.95786	28,663
GA	93.3426	990.1028	97.7	95.3	0.95935	1318
PSO	100	1000	97.8	95.6	0.96489	3506

3.4 多分类预测结果混淆矩阵可视化

将一阶差分-SFLA-PSO-SVM模型的预测结果以混淆矩阵的形式表示（图6）。其中，混淆矩阵主对角线上的绿色方框表明了特级、一级、二级和低质量牛乳预测正确的样本数和在总样本中所占的比例，红褐色方框则表明4类牛乳预测错误的样本数和在总样本中所占的比例，下、右的深灰色矩形框分别表示对应样本属性预测召回率和精准率。召回率即为正确预测为特级牛乳占全部实际为特级牛乳的比例、正确预测为一级牛乳占全部实际为一级牛乳的比例、正确预测为二级牛乳占全部实际为二级牛乳的比例以及正确预测为低质量牛乳占全部实际为低质量牛乳的比例。精准率即为正确预测为特级牛乳占全部预测为特级牛乳的比例、正确预测为一级牛乳占全部预测为一级牛乳的比例、正确预测为二级牛乳占全部预测为二级牛乳的比例以及正确预测为低质量牛乳占全部预测为低质量牛乳的比例。

由图6可知，测试集的964个样本中，特级、一级、二级和低质量牛乳的召回率分别为97.9%、94.8%、92.5%和96.0%，精准率分别为95.5%、95.5%、92.0%和98.1%，误判数量分别为6、13、12和11个。蓝色方框为模型预测准确率，为95.6%。

4 结 论

本研究以河北省9个牧场的3216份荷斯坦牛牛乳样本为研究对象，分别测定牛乳中的脂肪、蛋白质含量和体细胞数量并采集中红外光谱，构建了牛乳收购分级模型。主要结论如下：

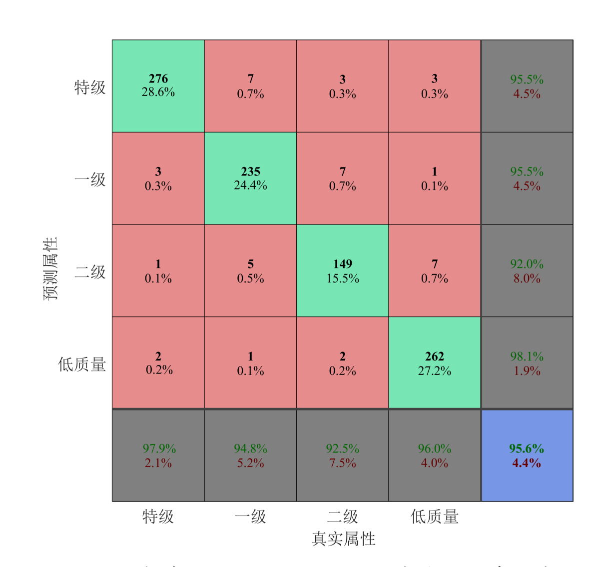


图6　一阶差分-SFLA-PSO-SVM模型的混淆矩阵

Fig. 6 Confusion matrix of first order differential-SFLA-PSO-SVM

（1）对特级、一级、二级和低质量牛乳的原始光谱和平均光谱进行分析并去除噪声波段和无贡献波段后，选择925～1597 cm^{-1}和1712～3024 cm^{-1}的敏感波段作为全光谱用于后续建模。

（2）对全光谱进行预处理后，为了剔除光谱冗余信息，克服维数灾难，结合CARS算法和SFLA算法进行特征变量筛选。结果表明，当利用CARS算法筛选特征变量时，一阶导数为最佳预处理算法，当利用SFLA算法筛选特征变量时，一阶差分为最佳预处理算法，SFLA算法总体上要优于CARS算法。最终选择一阶差分-SFLA-PSO-SVM模型为牛乳收购分级的最佳模型，训练集准确率、测试集准确率和AUC分别为97.8%、95.6%和0.96489。

（3）对比了GS、GA和PSO三种参数寻优算法的训练时间，结果表明GS的训练时间远长于POS和GA算法。

参考文献：

[1] 朱海明, 程启方. 瑞典牛奶检测分级付款系统简介[J]. 中国奶牛, 1997(4): 52-54.
ZHU H, CHENG Q. Brief introduction of Swedish milk testing grading payment system[J]. China Dairy Cattle, 1997(4): 52-54.

[2] 史慧茹, 姜瞻梅, 田波. 牛乳体细胞数的检测方法[J]. 畜牧与饲料科学, 2008(2): 86-88.
SHI H, JIANG Z, TIAN B. Method for detecting somatic cell count in bovine milk[J]. Animal Husbandry and Feed Science, 2008(2): 86-88.

[3] 陈贺, 王帅, 陈红玲. 乌鲁木齐地区生鲜牛乳质量分级研究[J]. 农村科技, 2017(8): 60-62.
CHEN H, WANG S, CHEN H. Study on the quality classification of fresh milk in Urumqi area[J]. Rural Science & Technology, 2017(8): 60-62.

[4] SMITH K L. Standards for somatic cells in milk: Physiologicaland regulatory[J]. IDF Mastitis Newslett, 1995, 144 (21): 7-9.

[5] KOLDWIJ E, EMANWLSON U. Relation of milk production lossto milk somatic cell count[J]. ACTA Vet Scand, 1999, 40: 47-56.

[6] GONDIM C, JUNQUEIRA R G, VITORINO C D S S, et al. Detection of several common adulterants in raw milk by MID-infrared spectroscopy and one-class and multi-class multivariate strategies[J]. Food Chemistry, 2017, 230: 68-75.

[7] TOFFANIN, V, PENASA, M, MCPARLAND, S, et al. Genetic parameters for milk mineral content and acidity predicted by mid-infrared spectroscopy in Holstein-Friesian cows[J]. Animal, 2015, 9(5): 775-780.

[8] SOYEURT H, DEHARENG F, GENGLER N, et al. Mid-infrared prediction of bovine milk fatty acids across multiple breeds, production systems, and countries[J]. Journal of Dairy Science, 2011, 94(4): 1657-1667.

[9] 李巧玲, 刘峰, 宋思远, 等. 中红外光谱法快速测定牛奶中非蛋白氮类物质[J]. 食品工业科技, 2014, 35 (22): 73-75, 80.
LI Q, LIU F, SONG S, et al. Fast determination of non-protein nitrogen content in milk based on mid-infrared spectroscopy method[J]. Science and Technology of Food Industry, 2014, 35(22): 73-75, 80.

[10] 吴珽, 梁龙, 朱华, 等. 海南制浆树种中主要成分的近红外分析与模型优化[J]. 光谱学与光谱分析, 2021, 41(5): 1404-1409.
WU T, LIANG L, ZHU H, et al. Near-infrared analysis and models optimization of main components in Pulpwood of Hainan province[J]. Spectroscopy and Spectral Analysis, 2021, 41(5): 1404-1409.

[11] 花晨芝, 赵凌, 宋建军, 等. 粒子群算法选择特征波长在紫外光谱检测COD中的研究[J]. 西华师范大学学报(自然科学版), 2019, 40(1): 81-85.
HUA C, ZHAO L, SONG J, et al. Selection of wavelength for UV-visible spectroscopy based on BLS com-

bined with PSO[J]. Journal of China West Normal University (Natural Sciences), 2019, 40(1): 81-85.

[12] 石吉勇, 邹小波, 王开亮, 等. 模拟退火算法用于食醋总酸含量近红外光谱模型的波数点优选[J]. 食品科学, 2011, 32(10): 120-123.
SHI J, ZOU X, WANG K, et al. Simulated annealing algorithm based wavenumber selection for total acid content analysis in vinegar by near infrared spectroscopy[J]. Food Science, 2011, 32(10): 120-123.

[13] 刘冬阳, 孙晓荣, 刘翠玲, 等. 拉曼光谱结合模拟退火的小麦粉灰分含量检测[J]. 中国粮油学报, 2019, 34(5): 128-133.
LIU D, SUN X, LIU C, et al. Detection of ash control of wheat flour based on Raman spectroscopy combined with simulated annealing[J]. Journal of the Chinese Cereals and Oils Association, 2019, 34(5): 128-133.

[14] 周孟然, 孙磊, 卞凯, 等. iPLS波段筛选方法在食用油品上快速检测研究[J]. 激光杂志, 2020, 41(7): 13-17.
ZHOU M, SUN L, BIAN K, et al. Band screening of iPLS for laser-induced fluorescence spectrum of edible oil[J]. Laser Journal, 2020, 41(7): 13-17.

[15] 张悉彦, 叶沁, 刘晓颖, 等. 傅里叶变换衰减全反射红外光谱结合向前区间偏最小二乘法快速测定食用油中总极性化合物[J]. 浙江农业科学, 2019, 60(6): 1003-1007.
ZHANG Z, YE Q, LIU X, et al. Fourier transform attenuated total reflection infrared spectroscopy combined with forward interval partial least squares method for rapid determination of total polar compounds in edible oil[J]. Journal of Zhejiang Agricultural Sciences, 2019, 60(6): 1003-1007.

[16] 王拓, 戴连奎, 马万武. 拉曼光谱结合后向间隔偏最小二乘法用于调和汽油辛烷值定量分析[J]. 分析化学, 2018, 46(4): 623-629.
WANG T, DAI L, MA W. Quantitative analysis of blended gasoline octane number using Raman spectroscopy with backward interval partial least squares method[J]. Chinese Journal of Analytical Chemistry, 2018, 46(4): 623-629.

[17] 史智佳, 李鹏飞, 吕玉, 等. 移动窗口偏最小二乘法优选猪油丙二醛近红外光谱波段[J]. 中国食品学报, 2014, 14(11): 207-213.
SHI Z, LI P, LYU Y, et al. Region optimization in FT-NIR spectroscopy for determination of MDA in lard with moving window partial least squares[J]. Journal of Chinese Institute of Food Science and Technology, 2014, 14(11): 207-213.

[18] 许良, 闫亮亮, 塞击拉呼, 等. 近红外光谱结合可移动窗口偏最小二乘法对克霉唑粉末药品的定量分析[J]. 计算机与应用化学, 2016, 33(4): 415-418.
XU L, YAN L, SAIJLAHU, et al. Quantitative analysis of Clotrimazole powder drugs by using moving window partial least square method combined with near-infrared spectroscopy[J]. Computers and Applied Chemistry, 2016, 33(4): 415-418.

[19] 李庆旭, 王巧华, 马美湖, 等. 基于可见/近红外光谱和深度学习的早期鸭胚雌雄信息无损检测[J]. 光谱学与光谱分析, 2021, 41(6): 1800-1805.
LI Q, WANG Q, MA M, et al. Non-destructive detection of male and female information of early duck embryos based on visible/near infrared spectroscopy and deep learning[J]. Spectroscopy and Spectral Analysis, 2021, 41(6): 1800-1805.

[20] 付丹丹, 王巧华, 高升, 等. 不同品种鸡蛋贮期S-卵白蛋白含量分析及其可见/近红外光谱无损检测模型研究[J]. 分析化学, 2020, 48(2): 289-297.
FU D, WANG Q, GAO S, et al. Analysis of S-Ovalbumin content of different varieties of eggs during storage and its nondestructive testing model by visible-near infrared spectroscopy[J]. Chinese Journal of Analytical Chemistry, 2020, 48(2): 289-297.

[21] 韩毅, 蔡建湖, 周根贵, 等. 随机蛙跳算法的研究进展[J]. 计算机科学, 2010, 37(7): 16-19.
HAN Y, CAI J, ZHOU G, et al. Advances in shuffled frog leaping algorithm[J]. Computer Science, 2010, 37(7): 16-19.

[22] 孙晶京, 杨武德, 冯美臣, 等. 基于随机蛙跳和支持向量机的冬小麦叶面积指数估算[J]. 山西农业大学学报(自然科学版), 2020, 40(5): 120-128.
SUN J, YANG W, FENG M, et al. Estimation of winter wheat leaf area index based on random leapfrog and support vector regression approach[J]. Journal of Shanxi Agricultural University (Natural Science Edition), 2020, 40(5): 120-128.

[23] 王巧华, 梅璐, 马美湖, 等. 利用机器视觉与近红外光谱技术的皮蛋无损检测与分级[J]. 农业工程学报, 2019, 35(24): 314-321..
WANG Q, MEI L, MA M, et al.Nondestructive testing and grading of preserved duck eggs based on machine vision and near-infrared spectroscopy[J]. Transactions of the CSAE, 2020, 40(5): 120-128.

[24] 黄平捷, 李宇涵, 俞巧君, 等. 基于SPA和多分类SVM的紫外-可见光光谱饮用水有机污染物判别方法研究[J]. 光谱学与光谱分析, 2020, 40(7): 2267-2272.
HUANG P, LI Y, YU Q, et al. Classify of organic contaminants in water distribution systems developed by SPA and multi-classification SVM using UV-VIS spectroscopy[J]. Spectroscopy and Spectral Analysis, 2020, 40(7): 2267-2272.

[25] Vapnik V N. An overview of statistical learning theory[J]. IEEE Transactions on Neural Networks, 1999, 10 (10): 988-999.

[26] Burges C J C. A Tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery. 1998, 2(2): 121-167.

[27] BONFATTI V, MARTINO G D, CARNIER P. Effectiveness of mid-infrared spectroscopy for the prediction of detailed protein composition and contents of protein genetic variants of individual milk of Simmental cows[J]. Journal of Dairy Science, 2010, 94(12): 5776-5785.

[28] 代芬, 邱泽源, 邱倩, 等. 基于拉曼光谱和自荧光光谱的柑橘黄龙病快速检测方法[J]. 智慧农业, 2019, 1 (3): 77-86.
DAI F, QIU Z, QIU Q, et al. Rapid detection of citrus Huanglongbing using Raman spectroscopy and autofluorescence spectroscopy[J]. Smart Agriculture, 2019, 1(3): 77-86.

[29] 胡翼然, 李杰庆, 刘鸿高, 等. 基于支持向量机对云南常见野生食用牛肝菌中红外光谱的种类鉴别[J]. 食品科学, 2021, 42(8): 248-256.
HU Y, LI J, LIU H, et al. Species identification of common wild edible bolete in Yunnan by Fourier transform mid-infrared spectroscopy coupled with support vector machine[J]. Food Science, 2021, 42(8): 248-256.

Construction of Milk Purchase Classification Model Based on Shuffled Frog Leaping Algorithm and Support Vector Machine

XIAO Shijie[1], WANG Qiaohua[1,2*], LI Chunfang[3,4], ZHAO Limei[4], LIU Xinya[4], LU Shiyu[4], ZHANG Shujun[3*]

(1. *College of Engineering*, *Huazhong Agricultural University*, *Wuhan* 430070, *China*; 2. *Key Laboratory of Agricultural Equipment in the Mid-Lower Reaches of the Yangze River*, *Ministry of Agriculture and Rural Affairs*, *Wuhan* 430070, *China*; 3. *Key Laboratory of Animal Breeding and Reproduction of Ministry of Education*, *Huazhong Agricultural University*, *Wuhan* 430070, *China*; 4. *Hebei Animal Husbandry Association*, *Shijiazhuang* 050031, *China*)

Abstract: Protein, fat and somatic cells are three important reference indicators in milk purchase, which determine the quality and price of milk. The traditional chemical analysis methods of these indexes are time-consuming and pollute the environment, while the mid-infrared spectrum has the advantages of fast, non-destructive and simple operation. In order to realize the rapid classification of milk quality and improve the production efficiency of dairy enterprises, 3216 Holstein milk samples were chosen as the research objects and mid-infrared spectroscopy technology was applied to realize the detection and classification of 4 different quality milks during the purchase process. The spectrum was preprocessed by using the first derivative and the first difference, and combined with the algorithm competitive adaptive reweighted sampling (CARS) and the shuffled frog leaping algorithm (SFLA), the effective characteristic variables that could represent different milks were selected, and the SVM model was established. Among them, the penalty parameter c and the kernel function parameter g which were the key parameters of the SVM model were optimized by using the grid search method (GS), genetic algorithm (GA) and particle swarm algorithm (PSO). The training time of GS, GA and PSO algorithms were compared, the results showed that the training time of GS was much longer than that of GA and PSO algorithms.The SFLA algorithm was generally better than the CARS algorithm, and the PSO optimized the SVM model the best. After the first-order difference preprocessing, the PSO-SVM established by using the SFLA algorithm to filter the characteristic variables, the accuracy of the training set, the accuracy of the test set and the AUC were 97.8%, 95.6% and 0.96489, respectively. This model has a high accuracy rate and has practical application value in the milk industry.

Key words: mid-infrared spectrum; milk; purchase classification; shuffled frog leaping algorithm; support vector machine